Nonparametric Definition of the Representativeness of a Sample—with Tables

By MILTON SOBEL and MARILYN J. HUYETT

(Manuscript received October 4, 1957)

The problem is to determine how large a random sample is needed in order to attain a preassigned probability $P^*(\frac{1}{2} \leqq P^* < 1)$ that the sample will possess a certain amount (or degree) of representativeness of the true unknown (cumulative) distribution F under study. The definition of representativeness involves two preassigned constants k and $\beta^*(k \geqq 2$ is an integer). For example, for $k = 2$ and any $\beta^*(0 < \beta^* \leqq \frac{1}{2})$ the sample is defined to be representative if the proportion of the total sample size falling on each side of the population median differs from $\frac{1}{2}$ by at most β^*. In this case the degree of representativeness is defined as $d_g^* = 1 - 2\beta^*$.

This idea can be extended to any number k of disjoint, exhaustive cells equi-probable under F; tables and graphs are given for finite and infinite populations for selected values of k, β^* and P^*. The definition is also extended to cases in which the experimenter is particularly interested in parts of F which are not equi-probable and/or parts of F which do not exhaust the whole sample space; tables and graphs accompany each application.

These results are non-parametric, i.e., if the prescribed sample size is used then the experimenter's requirements for representativeness will be satisfied whatever the unknown distribution. Derivations of exact and approximate formulae used in computing tables are given in the Appendices.

I. INTRODUCTION

This paper deals with the problem of determining how large a random sample is needed in order to guarantee with preassigned probability P^* that the sample will have a specified amount (or a specified degree) of representativeness of the true, unknown (cumulative) distribution F under study. No à priori information is given about F and no assumptions are made about the form of F. The solution given is nonparametric (i.e., distribution-free) so that the results obtained and the tables and graphs

constructed are valid for any true underlying distribution. The case of a finite population as well as that of an infinite population is considered; in the latter case it is assumed only for ease of exposition that those percentiles of F which enter the discussion are uniquely defined and have probability zero under F. (This will, in particular, be the case when F has a density function without zero-stretches between points having positive density.)

A definition of representativeness (and also a degree of representativeness) is given with respect to those parts of F which are between certain percentiles which we denote by $F^{-1}(p_i)$, the values of p_i being preassigned. The *intervals* between these percentiles will be called *cells* and we shall only consider collections of *pairwise disjoint* cells. For example the experimenter may want to guarantee with probability at least $P^* = 0.90$ that between 40 per cent and 60 per cent of his sample will lie on *each* side of the population median. In this case we are interested in the part of F (or the cell) between $F^{-1}(0)$ and $F^{-1}(0.5)$ and also the part of F (or the cell) between $F^{-1}(0.5)$ and $F^{-1}(1)$. By the definitions below the common allowance β^* is 0.10 and the degree of representativeness d_g^* is 0.80 (or 80 per cent). Then we enter Table 1 (or I1) with $k = 2$, $P^* = 0.90$ and $\beta^* = 0.10$ and find that the smallest sample size needed to satisfy the experimenter's requirement for representativeness is $n = 60$. (It is instructive to note that the same solution would hold for any two disjoint, exhaustive *subsets* of the sample space having a common probability of $\frac{1}{2}$ under F. However, the cases in which we consider disjoint *cells* and, in particular, disjoint cells which start from *one end* or *both ends* of the distribution are of considerably more practical interest. The *cell* terminology will be used in the body of the paper while the *subset* terminology will be used in the appendices.)

In the above example the sample space is broken up into two disjoint, exhaustive cells which are equi-probable under F. This idea of representativeness can be extended to any number k of pairwise disjoint, exhaustive cells equi-probable under F and in the numerical work the values $k = 2, 3, 4, 5$ and 10 are considered. The idea of representativeness can also be used with cells that are not equi-probable and/or with cells that do not exhaust the whole sample space. As an example of the first type (cells not equi-probable) we might be concerned about whether a sample is large enough to be *simultaneously* representative of a single tail with preassigned probability $p < \frac{1}{2}$ under F *and* of its complement which has probability $(1 - p) > \frac{1}{2}$ under F. As an example of the second type (non-exhaustive cells) we might be concerned about whether a sample

is large enough to be representative of both tails (each having (say) a common preassigned probability $p < \frac{1}{2}$ under F), without any concern about the middle cell between the two tails. For each problem tables and graphs throughout this paper give the smallest required sample size for selected values of P^* and specified amounts (or specified degrees) of representativeness.

Assuming for the moment that the density of F is known and that all of its deciles are finite then we can plot an observed bar diagram (i.e., rectangles with different widths under the *dashed lines* in Fig. 1) and the true density on the same diagram as shown in Fig. 1 to illustrate the idea of a representative sample. By definition of a decile each of the vertical strips bounded above by the *curve* has an area (or probability under F) of 0.1. The observed sample is considered representative relative to this pattern of ten disjoint, exhaustive and equi-probable cells to within a common allowance β^* if *simultaneously* the areas of *all* vertical *rectangles* differ from the theoretical value of 0.1 by at most $\beta^*(0 < \beta^* \leqq 0.1)$. Then the degree d_g^* of representativeness as defined in Section III is equal to $1 - 10\beta^*$. We are interested in finding the smallest sample size needed to guarantee a probability of at least P^* that the above condition will hold in a sample drawn at random from F.

This problem is related to the well-known problem[1] of Kolmogorov-Smirnov since they both have the common purpose of determining the sample size required to obtain a representative sample. Since their definition of representativeness is different from the one treated here, it is difficult to make a proper comparison of the two procedures. Another remark on this comparison is made in Appendix IV.

II. DEFINITION OF REPRESENTATIVENESS

Let F denote the true unknown cumulative distribution and let F_n^* denote the observed sample distribution based on n observations. For any *given* k let $C_1, C_2, \cdots, C_k$ denote pairwise disjoint cells (not necessarily exhaustive or equi-probable under F) which are defined by certain percentiles. The cells $C_1, C_2, \cdots, C_k$ are not known but their probabilities under F are given positive numbers; let $F(C_i)$ denote the probability assigned to C_i by the distribution $F(i = 1, 2, \cdots, k)$. (We are using F and F_n^* as symbols for both point functions and probability measures which are set functions; clearly, the nature of the argument will prevent any confusion.) Let β_i^* denote specified positive numbers (which we shall call allowances) such that

$$0 < \beta_i^* \leqq F(C_i) \qquad (i = 1, 2, \cdots, k). \quad (1)$$

We shall be particularly interested in the special case $\beta_1^* = \beta_2^* = \cdots = \beta_k^* = \beta^*$ (say), whether or not the quantities $F(C_i)$ are all equal. Then a sample is defined to be representative relative to a fixed pattern of k disjoint cells $C_1, C_2, \cdots, C_k$ to within the allowances $\beta_1^*, \beta_2^* \cdots, \beta_k^*$, respectively, if we have *simultaneously*

$$| F_n^*(C_i) - F(C_i) | \leqq \beta_i^* \qquad (i = 1, 2, \cdots, k). \quad (2)$$

III. DEFINITION OF DEGREE OF REPRESENTATIVENESS

Although the quantities $\beta_i^*(i = 1, 2, \cdots, k)$ are basic to the idea of representativeness it may be useful, in a given problem, to combine them to define a measure of the *degree* of representativeness. We define

$$d_g^* = \left\{ \prod_{i=1}^{k} \left[1 - \frac{\beta_i^*}{F(C_i)} \right] \right\}^{1/k} \quad (3)$$

where the subscript g denotes the fact that d_g^* is a *geometric* mean. It follows from (1) that $0 \leqq d_g^* < 1$ and that d_g^* can take on all the values in this interval.

It should be noted that for any fixed set of values of $F(C_i)$ $(i = 1, 2, \cdots, k)$ if there is a common β^* then the right band member of (3) is a strictly decreasing function of β^* for $\beta^* \leqq \min F(C_i)$. Hence, if there is a common β^* the values of d_g^* and β^* uniquely determine each other. When this is the case we may be interested sometimes in specifying d_g^* (instead of β^*) and then using (3) to solve for the common β^*.

We shall say that a random sample is representative relative to a fixed pattern of k disjoint cells $C_1, C_2, \cdots, C_k$ to a degree d_g^* if for *the common* $\beta^* = \beta^*(d_g^*)$ satisfying (3) we bave

$$| F_n^*(C_i) - F(C_i) | \leqq \beta^* \qquad (i = 1, 2, \cdots, k). \quad (4)$$

It should be emphasized that the chief interest of this paper is in the concept of representativeness as formulated in Section II and that the present definition of the *degree* of representativeness is to be regarded as supplementary.

One possible criticism of the definition of d_g^* is tbat it may require a positive (and sometimes substantial) number of observations to attain a *zero* degree of representativeness (see, for example, the last and third from last columns in Table III). However, since the practical use of the concept of *degree* of representativeness is mainly for *large* values of d_g^* this objection is not serious.

It is possible also to define the *degree* of representativeness as an *arithmetic* mean d_a^* of the bracketed quantities in (3) but then for a common β^* and different $F(C_i)$, because of (1), the value of d_a^* is restricted to an interval $J \leqq d_a^* < 1$ where J is *positive* and depends on the values of the $F(C_i)$ $(i = 1, 2, \cdots, k)$. Clearly, if the $F(C_i)$ are all equal and there is a common β^* then $d_a^* = d_g^*$.

IV. CONSTRUCTION OF TABLES

The problem is to find the *smallest* sample size n such that the joint probability of all the inequalities (2) [or (4)] is at least equal to a specified value $P^* < 1$, i.e., such that

$$P\{ \mid F_n^*(C_i) - F(C_i) \mid \leqq \beta_i^*(i = 1, 2, \cdots, k)\} \geqq P^*. \tag{5}$$

The reader is cautioned that it does not necessarily follow that (5) holds for any integer greater than n; however, since F_n^* converges almost certainly to F (see page 20 of Reference 2), it follows that there exists in each case a smallest number $n' \geqq n$ such that (5) holds for *every integer* greater than or equal to n'. For example, with $k = 2$, a common $\beta^* = 0.20$ and $P^* = 0.75$ the condition (5) is satisfied for $n = 3$, for 6 and for any integer greater than or equal to $n' = 9$.

Since the cells C_i are pairwise disjoint and the values of $F(C_i)$ are given $(i = 1, 2, \cdots, k)$ the left member of (5) is determined for any particular sample size whatever the unknown distribution F. In the case of an infinite population we use the multinomial distribution with k or $k + 1$ disjoint cells depending on whether or not the k disjoint cells are exhaustive, i.e., on whether or not $\sum_{i=1}^{k} F(C_i) = 1$. For the case of two disjoint, exhaustive cells this clearly reduces to a problem of the binomial distribution which is closely related to the problem of finding confidence limits on a population percentile by the use of order statistics. Similarly in the case of a finite population we use the hypergeometric distribution with k or $k + 1$ categories depending on whether or not $\sum_{i=1}^{k} F(C_i) = 1$. The exact and approximate formulae for computing the left member of (5) are given in Appendices I and II, respectively. The approximate calculation involves several interesting geometrical digressions which are discussed in Appendix III.

Table I gives for $k = 2$ and selected values of β^* and P^* the required sample sizes n *and* n' and also the maximum drop in probability below the specified P^* for all sample sizes between n and n'. In the remaining tables only the values of n are given. Table II gives the required sample size for $k = 2$, $F(C_1) = p$, $F(C_2) = 1 - p$ for $p = 0.5, 0.2$ and 0.1 (for

TABLE I

Sample size required to attain a probability P^* that a sample will be simultaneously representative to within a common allowance β^* of two disjoint and exhaustive cells separated by the median for any true distribution.

In each set the first entry is the smallest sample size required to satisfy (4); the second entry is the smallest size required such that for *all* sample sizes at least as large, (4) is satisfied; the last entry is the maximum deviation in probability below P^* obtained for all sample sizes between the first two entries.

P^* \ β^*	0.01	0.05	0.10	0.15	0.20	0.25	0.40
0.50	1051 1199 (0.0264)	31 59 (0.1271)	5 14 (0.2266)	5 10 (0.1875)	2 5 (0.1250)	2 2 (0)	2 2 (0)
0.60	1700 1850 (0.0162)	60 79 (0.0704)	5 24 (0.3266)	5 10 (0.2875)	3 8 (0.2250)	3 3 (0)	3 3 (0)
0.70	2600 2750 (0.0124)	100 119 (0.0382)	20 29 (0.1049)	8 16 (0.2078)	3 8 (0.3250)	3 6 (0.0750)	3 3 (0)
0.75	3251 3399 (0.0077)	120 150 (0.0407)	25 39 (0.0769)	11 16 (0.1377)	3 9 (0.3750)	3 6 (0.1250)	3 3 (0)
0.80	4051 4199 (0.0058)	151 179 (0.0328)	35 44 (0.0430)	14 24 (0.0518)	9 12 (0.0266)	4 7 (0.0750)	4 4 (0)
0.85	5100 5250 (0.0052)	191 219 (0.0269)	45 54 (0.0434)	17 27 (0.0879)	10 15 (0.0766)	4 10 (0.1250)	4 4 (0)
0.90	6700 6850 (0.0029)	260 279 (0.0129)	60 74 (0.0299)	28 33 (0.0360)	13 18 (0.0796)	8 11 (0.0797)	5 5 (0)
0.95	9551 9699 (0.0012)	371 399 (0.0070)	90 99 (0.0114)	37 47 (0.0230)	20 28 (0.0284)	12 15 (0.0423)	6 6 (0)
0.99	16500 16650 (0.0003)	651 679 (0.0013)	160 169 (0.0022)	71 76 (0.0028)	39 42 (0.0015)	24 26 (0.0046)	8 12 (0.0017)

For $n \leqq 150$ the entries are all exact; for $n > 150$ the entries involve approximations. The pattern of increases and decreases of the probability as a function of n was also used to obtain the first two entries for large n.

selected values of β^* and P^*). Table III gives the required sample size for the case of k pairwise disjoint, exhaustive and equi-probable cells ($C_1, C_2, \cdots, C_k$) for $k = 2, 3, 4, 5$ and 10 (for selected values of β^* and P^*). Table IV gives the required sample size for $k = 2$, $F(C_1) = F(C_2) = p$ for $p = 0.2, 0.1$ and 0.05 (here the cells are disjoint and equi-probable but not exhaustive). Table V considers the same problem as in Table III and compares the required sample sizes for infinite populations, $N = \infty$, with those for finite populations of size N for $N = 60$, 120, 360. Tables VI and VII give illustrations of the error involved in using the approximations used in Tables IV and V, respectively, instead of an exact probability calculation.

Fig. 2 shows for selected values of P^* that the sample sizes in Table I and in the first portion of Table II can be "linearized" for large n on a log-log plot of n versus β^*. Figs. 3 and 4 show the same result for the last and middle portion of Table II, respectively.

Table II

Minimum sample size required to attain a probability of at least P^* that a sample will be simultaneously representative to within a common allowance β^* of two disjoint and exhaustive cells separated by the 100 pth percentile for any true distribution. (The degree of representativeness is then defined as $d_\sigma^* = \sqrt{\left(1 - \frac{\beta^*}{p}\right)\left(1 - \frac{\beta^*}{1-p}\right)}$.)

	50th Percentile (Median) ($p = 0.50$)					20th or 80th Percentile ($p = 0.20$ or 0.80)					10th or 90th Percentile ($p = 0.10$ or 0.90)		
P^* \ β^*	0.01	0.05	0.10	0.15	0.20	0.01	0.05	0.10	0.15	0.20	0.01	0.05	0.10
0.50	1,051	31	5	5	2†	662	12	7	6	1†	355	14	1†
0.60	1,700	60	5	5	3†	1,062	32	7	6	1†	500	14	1†
0.70	2,600	100	20	8	3†	1,662	52	10	9	1†	900	20	1†
0.75	3,251	120	25	11	3†	2,062	72	10	9	1†	1,100	40	1†
0.80	4,051	151	35	14	9	2,562	92	20	12	1†	1,400	40	1†
0.85	5,100	191	45	17	10	3,262	120	27	12	3†	1,800	60	1†
0.90	6,700	251	60	28	13	4,262	160	37	15	5	2,355	80	1†
0.95	9,551	371	90	37	20	6,100	232	50	20	10	3,400	120	10
0.99	16,500	651	160	71	39	10,562	420	100	40	20	5,900	220	15

For $n \leq 150$ the entries are all exact; for $n > 150$ the entries are based on approximations together with a knowledge of the monotonicity pattern of the probability of representativeness as a function of n.

† Small entries for certain pairs (β^*, P^*) indicate a condition too weak for practical usage.

TABLE III

Minimum sample size required to attain a probability of at least P^* that a sample will be simultaneously representative to within a common allowance β^* of k equi-probable disjoint and exhaustive cells for any true distribution. (The degree of representativeness is then defined as $d_g^* = 1 - k\beta^*$).

	$k = 2$			$k = 3$			$k = 4$			$k = 5$			$k = 10$	
P^* \ β^*	0.05	0.10	0.20	0.05	0.10	0.20	0.05	0.10	0.20	0.05	0.10	0.20	0.05	0.10
0.50	31	5	2	102	21	6	120	26	9	120	30	5	100	20
0.60	60	5	3	141	30	6	140	38	9	140	30	5	100	20
0.70	100	20	3	180	47	12	180	43	12	180	40	5	120	30
0.75	120	25	3	222	51	14	200	52	14	200	50	10	120	30
0.80	151	35	9	240	60	15	240	60	14	220	50	10	140	30
0.85	191	45	10	300	72	15	280	66	16	240	60	15	160	30
0.90	251	60	13	360	90	21	320	80	18	280	70	15	160	40
0.95	371	90	20	480	120	29	400	100	27	360	90	23	200	50
0.99	651	160	39	741	180	45	600	146	38	500	120	35	260	60

For $k \geqq 3$ probabilities were computed exactly only for $n \leqq (200/k)$; for $n > (200/k)$ the approximation in Appendix 2 was used together with a knowledge of the monotonicity pattern of the probability of representativeness as a function of n.

TABLE IV

Minimum sample size required to attain a probability of at least P^* that a sample will be simultaneously representative to within a common allowance β^* of any two disjoint equi-probable cells defined by percentiles and having a common probability p under the true, unknown distribution. (The degree of representativeness is then defined as $d_g^* = 1 - \beta^*/p$.)

Application	Below 20th and Above 80th Percentiles ($p = 0.20$)			Below 10th and Above 90th Percentiles ($p = 0.10$)			Below 5th and Above 95th Percentiles ($p = 0.05$)	
P^* \ β^*	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05
0.50	1,700	52	10	900	20	1†	450	1†
0.60	2,262	72	10	1,255	40	1†	600	1†
0.70	3,000	112	20	1,655	54	1†	850	1†
0.75	3,500	132	30	1,955	60	1†	1,000	1†
0.80	4,100	152	30	2,300	80	1†	1,150	1†
0.85	4,900	180	40	2,700	100	10	1,400	1†
0.90	6,000	232	50	3,355	120	20	1,750	1†
0.95	7,900	300	70	4,455	160	35	2,250	80
0.99	12,562	492	120	7,000	274	65	3,650	130
Another Application	Between 30th and 50th percentiles and between 50th and 70th percentiles			Between 40th and 50th percentiles and between 50th and 60th percentiles			Between 45th and 50th percentiles and between 50th and 55th percentiles	

For $n \leqq 40$ the entires are exact; for $n > 40$ normal approximation theory was used.

† Small entires for certain pairs (β^*, P^*) indicate a condition too weak for practical usage

Table V

Minimum sample size required to attain a probability of at least P^* that a sample from a population of size N will be simultaneously representative to within a common allowance β^* of k equi-probable disjoint and exhaustive cells for any true population. (The degree of representativeness is then defined as $d_g^* = 1 - k\beta^*$).

The four entries in each set below correspond to $N = 60, 120, 360, \infty$, respectively.

	$k = 2$			$k = 3$			$k = 4$			$k = 5$			$k = 10$	
P^* \ β^*	0.05	0.10	0.20	0.05	0.10	0.20	0.05	0.10	0.20	0.05	0.10	0.20	0.05	0.10
0.50	20 20 20 31	5 5 5 5	2 2 2 2	40 55 81 102	19 21 21 21	6 6 6 6	40 60 80 120	20 20 20 26	7 7 7 7	40 60 80 120	20 20 24 30	3 5 5 5	34 54 74 100	10 15 15 20
0.75	40 60 91 120	15 20 25 25	3 3 3 3	47 76 136 222	28 37 49 51	12 14 14 15	47 74 130 200	26 38 40 52	12 12 14 14	45 72 120 200	27 30 40 50	8 8 10 10	40 60 94 120	20 25 25 30
0.85	51 71 120 191	25 30 40 45	9 10 10 10	53 84 162 300	30 49 60 72	14 15 15 15	50 80 150 280	32 40 58 66	14 14 16 16	49 80 152 240	30 40 50 60	10 10 13 15	40 60 100 160	20 25 30 30
0.90	51 80 151 251	30 40 50 60	10 13 13 13	54 93 180 360	37 51 72 90	15 19 21 21	50 90 170 320	38 46 60 80	16 16 18 20	51 80 160 280	30 40 60 70	13 13 15 15	40 74 114 160	25 25 35 40
0.95	51 91 180 371	35 50 70 90	16 19 20 20	54 94 201 480	42 60 88 120	21 25 27 30	50 90 190 400	38 58 80 100	18 20 25 27	52 92 180 360	37 50 70 90	15 15 18 20	47 74 120 200	25 30 40 50
0.99	60 100 231 651	45 70 110 160	23 30 36 39	55 102 240 741	48 72 120 180	27 30 42 45	57 100 220 600	43 66 100 146	25 29 34 37	53 98 212 500	40 60 90 120	20 23 25 30	49 80 154 260	30 40 50 60

For finite populations all entries with $n \leqq 2/\beta^*$ are based on exact computations; the entries with $n > 2/\beta^*$ are based on the approximation in equation (A17) of Appendix II. Another simpler approximation is given in equation (A18) of Appendix II.

Table VI

Comparison between the exact value of and the normal approximation to the joint probability that in a sample of size n from an infinite population the number of observations falling in each of two tails with common probability p is between $n(p - \beta^*)$ and $n(p + \beta^*)$, inclusive.

		$p = 0.10$, $\beta^* = 0.05$	$p = 0.20$, $\beta^* = 0.05$	$p = 0.20$, $\beta^* = 0.10$
$n = 10$	Normal Approx.	0.1628	0.0973	0.5910
	Exact	0.1510	0.0941	0.6014
	Error	+0.0118	+0.0032	−0.0104
$n = 20$	Normal Approx.	0.5432	0.3654	0.7075
	Exact	0.5566	0.3648	0.7171
	Error	−0.0134	+0.0006	−0.0096
$n = 40$	Normal Approx.	0.6608	0.4655	0.8574
	Exact	0.6731	0.4669	0.8736
	Error	−0.0123	−0.0014	−0.0162

Table VII

Comparison between the exact value of and the normal approximation to the joint probability that in a sample of size n from a population of size N the number of observations falling in each of k equi-probable cells is between $n\left(\frac{1}{k} - \frac{1}{20}\right)$ and $n\left(\frac{1}{k} + \frac{1}{20}\right)$, inclusive.

$N = \infty$ (Infinite Population)

		$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 10$
$n = 20$	Normal Approx.	0.4977	0.1166	0.1600	0.1172	0.0698
	Exact	0.4966	0.1145	0.1618	0.0955	0.0669
	Error	+0.0011	+0.0021	−0.0018	+0.0217	+0.0029
$n = 40$	Normal Approx.	0.5708	0.2196	0.2388	0.1962	0.1775
	Exact	0.5704	0.2181	0.2363	0.1904	0.1478
	Error	+0.0004	+0.0015	+0.0025	+0.0058	+0.0297
$n = 60$	Normal Approx.	0.6338	0.3974	0.3230	0.2876	0.3325
	Exact	0.6338	0.3982	0.3174	0.2979	*
	Error	0.0000	−0.0008	+0.0056	−0.0103	*

$N = 120$ (Finite Population)

		$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 10$
$n = 20$	Normal Approx.	0.5357	0.1397	0.1984	0.1550	0.1092
	Exact	0.5368	0.1359	0.1801	0.1547	0.1011
	Error	−0.0011	+0.0038	+0.0183	+0.0003	+0.0081
$n = 40$	Normal Approx.	0.6651	0.2822	0.3705	0.3413	0.4291
	Exact	0.6670	0.3084	0.3679	0.3313	0.3357
	Error	−0.0019	−0.0262	+0.0026	+0.0100	+0.0934
$n = 60$	Normal Approx.	0.7969	0.6338	0.6115	0.6228	0.8507
	Exact	0.7989	0.6104	0.6003	0.5972	*
	Error	−0.0020	+0.0234	+0.0112	+0.0256	*

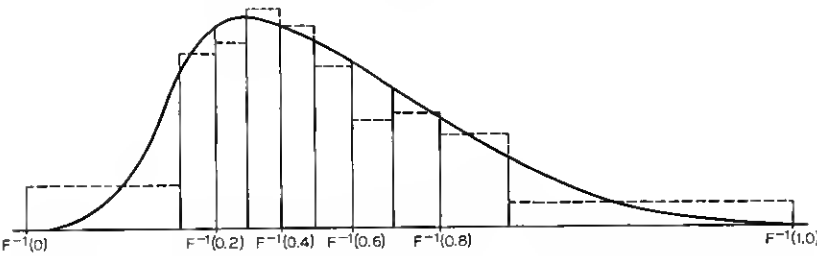

Fig. 1 — Pictorial diagram of representativeness using deciles ($k = 10$).

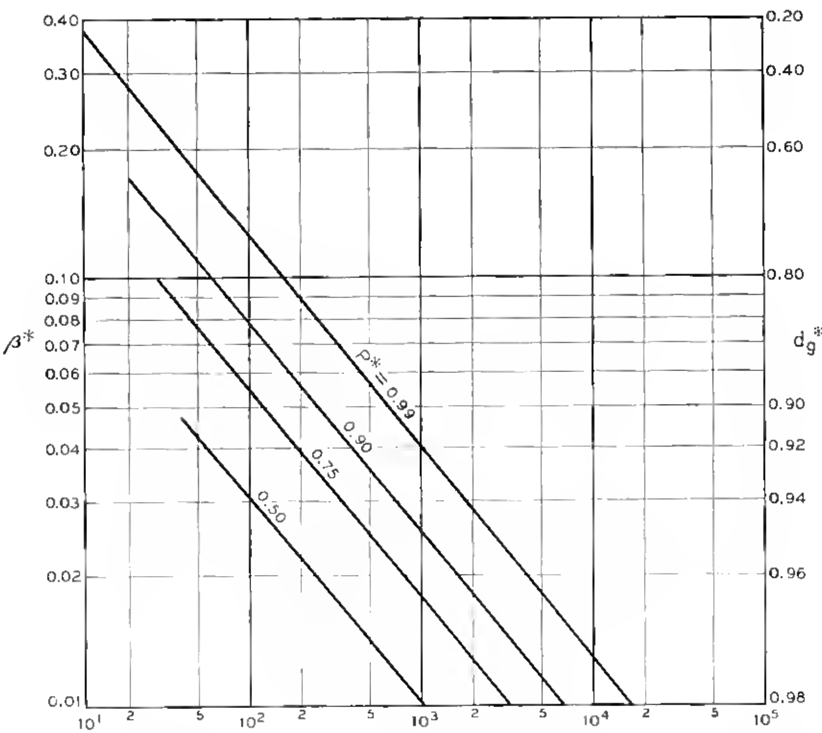

Fig. 2 — Minimum sample size n required to attain a probability of at least P^* that a sample is simultaneously representative to within a common allowance β^* of two disjoint and exhaustive cells each having probability $p = \frac{1}{2}$ under the true unknown distribution. (The degree of representativeness is $d_g^* = 1 - 2\beta^*$.)

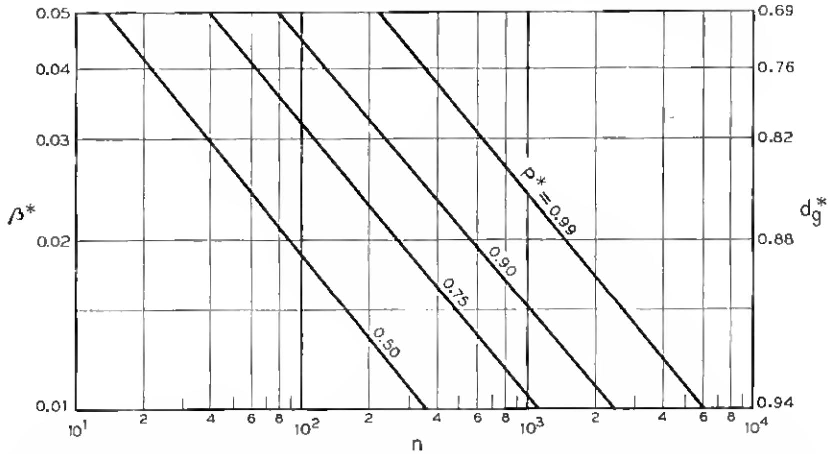

Fig. 3 — Minimum sample size n required to attain a probability of at least P^* that a sample is simultaneously representative to within a common allowance β^* of the two disjoint, exhaustive cells separated by the 10th (or the 90th) percentile for any true distribution. [The degree of representativeness is $d_g^* = (\frac{10}{3}) \sqrt{(0.1 - \beta^*)(0.9 - \beta^*)}$.]

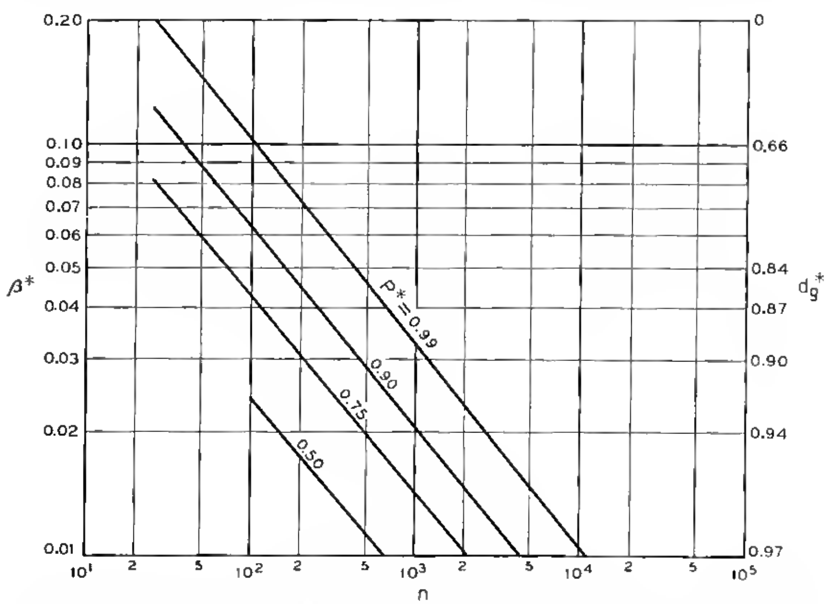

Fig. 4 — Minimum sample size n required to attain a probability of at least P^* that a sample is simultaneously representative to within a common allowance β^* of the two disjoint, exhaustive cells separated by the 20th (or the 80th) percentile for any true distribution. [The degree of representativeness is $d_g^* = (\frac{5}{2}) \sqrt{(0.2 - \beta^*)(0.8 - \beta^*)}$.]

V. EMPIRICALLY OBSERVED MONOTONICITIES

It is interesting to note in Table III that for fixed β^* and increasing k the sample size n required is *not* monotonic but appears to reach a maximum and then decrease. As a result of this it becomes possible to to speak of the sample size n required for a sample to be representative for any specified β^* regardless of the number k of pairwise disjoint, exhaustive, equi-probable cells considered, provided only that $k \leqq 1/\beta^*$. For example, for $\beta^* = 0.1$ it appears likely from Table III that 90 observations would be sufficient to have a confidence of at least $P^* = 0.90$ that the sample is representative in the sense of (2) for *any one value* of $k(k = 1, 2, \cdots, 10)$.

Table VIII, some of whose entries are taken from Table III, shows *numerically* that *for fixed* d_g^* the required sample size is a monotonically non-decreasing function not only of P^* *but also of* k; *for fixed* β^*. Table III shows numerically that only the monotonicity with P^* holds. The former result is again shown in Figs. 5 and 6 which also emphasize the possibilities of interpolation on k.

The above monotonicities and lack of monotonicities have not been demonstrated mathematically.

TABLE VIII

Minimum sample size required to attain a probability of at least P^* that a sample will be simultaneously representative to a degree $d_g^* = 1 - k\beta^*$ of k equi-probable disjoint and exhaustive cells for any true distribution.

P^*	$d_g^* = 0.80$			$d_g^* = 0.90$		
	$k = 2$	$k = 4$	$k = 10$	$k = 2$	$k = 5$	$k = 10$
0.50	5	120	600	31	800	2500
0.60	5	140	700	60	950	2800
0.70	20	180	800	100	1150	3200
0.75	25	200	850	120	1250	3400
0.80	35	240	900	151	1400	3700
0.85	45	280	1000	191	1600	4000
0.90	60	320	1100	251	1850	4400
0.95	90	400	1250	371	2250	5100
0.99	160	600	1650	651	3150	6600

In comparing results for a *fixed degree* d_g^* it should be noted that the sample size appears to be a monotonically non-decreasing function of P^* *and* also of k; for a *fixed common* allowance β^* only the monotonicity with P^* holds as is evident in Table II. The remarks at the bottom of Table III apply here also.

VI. CONFIDENCE BANDS—INFINITE POPULATION CASE

The experimenter will usually be interested in the confidence statement that the above formulation allows him to make *after the observations are taken*. Suppose, for example, that he was interested in representativeness in each of $k = 10$ pairwise disjoint, exhaustive and equi-probable cells and that he specified $\beta^* = 0.02$ (so that $d_g^* = 0.80$) and $P^* = 0.85$ and that he has taken 1,000 observations in accordance with Table VIII.

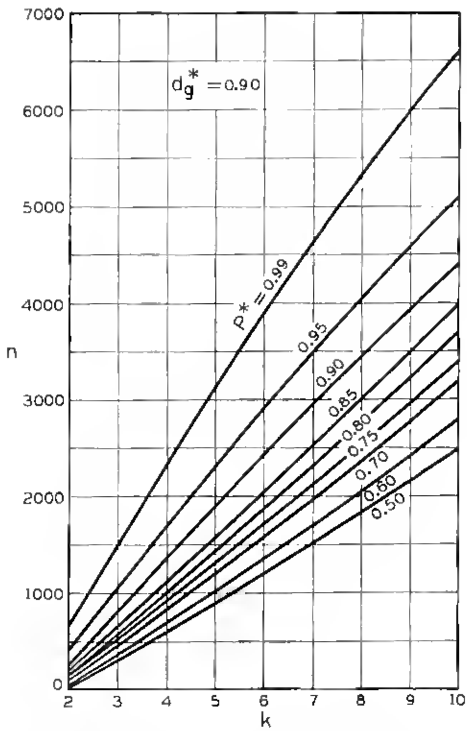

Fig. 5 — Minimum sample size n required to attain a probability of at least P^* that a sample will be simultaneously representative to a degree $d_g^* = 0.90$ of k equi-probable, disjoint and exhaustive cells for any true distribution. The common allowance β^* is given by $\beta^* = (1 - d_g^*)/k = 0.10/k$.

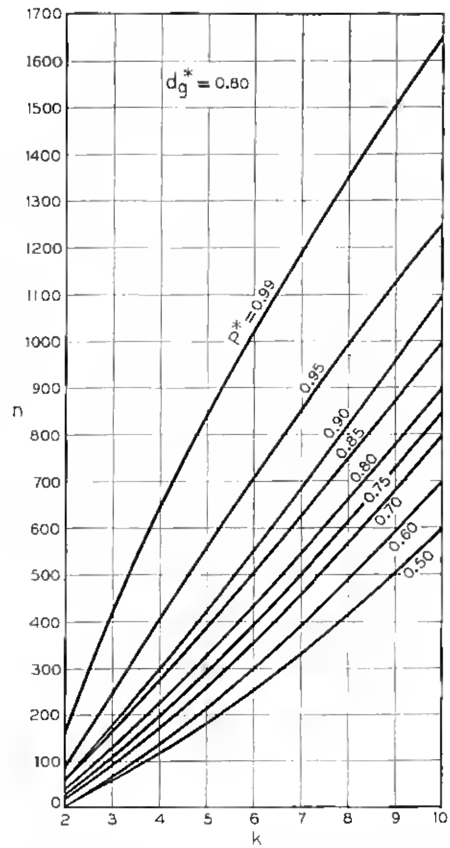

Fig. 6 — Minimum sample size n required to attain a probability of at least P^* that a sample will be simultaneously representative to a degree $d_g^* = 0.80$ of k equi-probable, disjoint and exhaustive cells for any true distribution. The common allowance β^* is given by $\beta^* = (1 - d_g^*)/k = 0.20/k$.

He can then make a number of confidence statements about the population deciles $F^{-1}(0.1)$, $F^{-1}(0.2)$, $\cdots$, $F^{-1}(0.9)$ (and also about $F^{-1}(0)$ and $F^{-1}(1)$ defined as the greatest lower bound of all x for which $F(x) > 0$ and the least upper bound of all x for which $F(x) < 1$, respectively). For example, if x_m denotes the mth (smallest) ordered observation, it follows from the condition of representativeness that we have *simultaneously* with joint confidence greater than P^* all of the inequalities

$$\left\{\begin{array}{lcl} -\infty & \leqq F^{-1}(0) & < x_1 \\ x_{80} & \leqq F^{-1}(0.1) & < x_{121} \\ x_{160} & \leqq F^{-1}(0.2) & < x_{241} \\ \cdots & \cdots & \cdots \\ x_{640} & \leqq F^{-1}(0.8) & < x_{961} \\ x_{720} & \leqq F^{-1}(0.9) & < \infty \\ x_{1000} & \leqq F^{-1}(1) & \leqq \infty \end{array}\right\} \text{ and } \left\{\begin{array}{lcl} x_{1000} & < F^{-1}(1) & \leqq \infty \\ x_{879} & < F^{-1}(0.9) & \leqq x_{920} \\ x_{759} & < F^{-1}(0.8) & \leqq x_{840} \\ \cdots & \cdots & \cdots \\ x_{39} & < F^{-1}(0.2) & \leqq x_{360} \\ -\infty & < F^{-1}(0.1) & \leqq x_{280} \\ -\infty & \leqq F^{-1}(0) & \leqq x_1 \end{array}\right\}. \quad (6)$$

For example, $F^{-1}(0.2)$ must be greater than or equal to x_{160} and less than x_{241} in the confidence statement since under the condition of representativeness all cells and, in particular, the last two cells on the left contain between 80 and 120 observations, inclusive.

The right hand set of inequalities are in reverse order since they are obtained by similar reasoning as the left hand set except that we start at the right end of the distribution and work backwards. If we keep only the stronger results in (6) for each decile and disregard the weaker ones, then we obtain eleven (finite or infinite) line segments as in Fig. 7. We can then state with joint confidence greater than P^* that the unknown distribution F has a (finite or infinite) point of contact with (or a saltus passing through) each of the line segments; the two end segments are

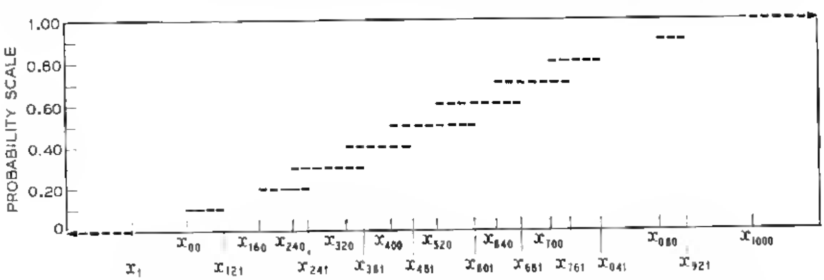

Fig. 7 — Confidence intervals for the deciles with joint confidence level $P^* = 0.85$ for $k = 10$, $n = 1000$ and $\beta^* = 0.02$ (which implies that $d_0^* = 0.80$).

actually half-lines and in these cases we must allow $+\infty$ and $-\infty$ as possible "points" of contact.

The above result then gives rise to two "staircases", as in the middle diagram of Fig. 8, such that any distribution contacting every line segment in Fig. 7 must everywhere lie between (or on the boundary of) the two "staircases". Hence we can state with confidence *greater than* P^* (see explanation below) that the two "staircases" form a confidence band on the unknown distribution.

If we keep k and P^* fixed and decrease β^* (or increase $d_g^* = 1 - k\beta^*$)

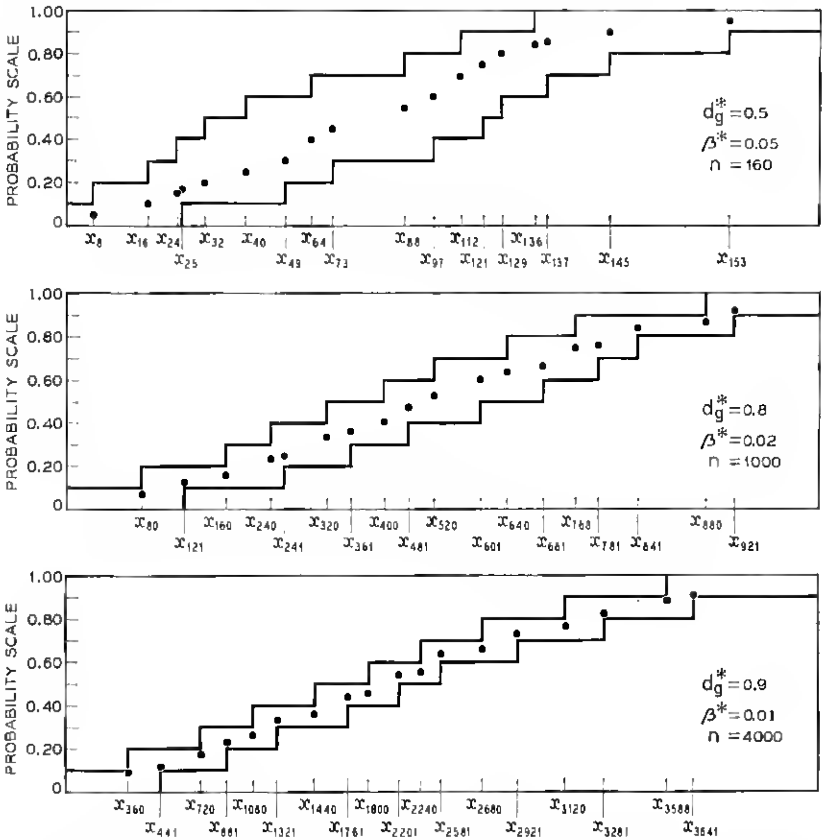

Fig. 8 — Confidence bands which include the true distribution function with confidence greater than $P^* = 0.85$ for $k = 10$ and $d_g^* = 0.5, 0.8, 0.9$. Small circles between the confidence bands represent ordinates of the sample distribution function. The three figures above were constructed with observations obtained from a table of random normal deviates (with different horizontal scaling applied in each case).

then the required sample size increases and the confidence band becomes narrower. This is illustrated in the three diagrams of Fig. 8.

It should be noted that the inequalities (6) are implied by but do not imply (i.e., they are not equivalent to) the condition of representativeness. Hence the confidence level associated with (6) is *greater than* the specified P^*. To illustrate this we note from (6) the stronger inequalities

$$x_{80} \leqq F^{-1}(0.1) < x_{121} \quad \text{and} \quad x_{160} \leqq F^{-1}(0.2) < x_{241} . \tag{7}$$

These inequalities (7) allow as few as 40 and as many as 161 observations between $F^{-1}(0.1)$ and $F^{-1}(0.2)$, including endpoints. On the other hand we have confidence P^*, under the condition of representativeness, that every such cell contains between 80 and 120 observations, inclusive. This shows that the confidence level associated with the confidence band is greater than the probability achieved for the representativeness of the sample.

This method of obtaining a confidence band for the unknown distribution would be more valuable if we could obtain a simple way of computing (or estimating more accurately) the actual confidence level attained. For example, with $k = 3$, $\beta^* = 0.10$ (so that $d_g^* = 0.70$) and $P^* = 0.60$ we obtain $n = 30$ from Table III, the probability achieved for representativeness is 0.6369 and the confidence level associated with the two "staircases" is 0.6825. The latter is obtained by using inequalities similar to (6) and computing the probability exactly with a multinomial distribution. The reader should note that the idea of a confidence band containing the true, unknown distribution is not the main theme of this paper but only an interesting by-product of the idea of the representativeness of the sample.

APPENDIX I

Exact Formulae — Finite and Infinite Populations

The concept of the representativeness of a sample can be applied to finite as well as infinite populations. Let N denote the total size of a finite population; conceptually we may regard the population as being partitioned into k subsets S_i of size $F(S_i)(i = 1, 2, \cdots, k)$. We shall assume that the sets S_i are pairwise disjoint and, to simplify the discussion, we also assume that the quantities $N_i = NF(S_i)(i = 1, 2, \cdots, k)$ are *positive integers*.

Let $x_i \geqq 0$ denote the random integral number of observations in the observed sample of size n which fall in the set $S_i(i = 1, 2, \cdots, k)$. If

the k sets S_i are exhaustive then

$$\textstyle\sum_{i=1}^{k} x_i = n \qquad \text{and} \qquad \sum_{i=1}^{k} N_i = N. \tag{A1}$$

We define for $i = 1, 2, \cdots, k$

$$c_i = n[F(S_i) - \beta_i^*] \quad \text{and} \quad d_i = n[F(S_i) + \beta_i^*], \tag{A2}$$

which are non-negative but need not be integers. Then for a finite population the probability corresponding to the left number of (5), using the hypergeometric distribution, is given *exactly* by

$$P_n^{(N)}[N_i, a_i, b_i \ (i = 1, 2, \cdots, k)] = \sum \prod_{i=1}^{k} \binom{N_i}{x_i} \Big/ \binom{N}{n} \tag{A3}$$

where $\binom{N}{n}$ is the usual binomial coefficient and the summation in (A3) is over all vectors $\vec{x} = \{x_1, x_2, \cdots, x_k\}$ for which

$$c_i \leqq x_i \leqq d_i \qquad (i = 1, 2, \cdots, k). \tag{A4}$$

If the k sets are *not* exhaustive then we define another set S_{k+1} which is the complement of the union of the k sets S_i and use (A3) with k replaced by $k + 1$ in (A1) and (A3) but *not* in (A4), i.e., no condition is applied to the $(k + 1)$th variable.

In the case of an infinite population we use the multinomial distribution. If the k sets S_i are exhaustive, then using (A2) and letting $p_i = F(S_i)(i = 1, 2, \cdots, k)$ the left hand member of (5) is given *exactly* by

$$P_n^{(\infty)}[p_i, \beta_i^* \ (i = 1, 2, \cdots, k)] = \sum \frac{n!}{\prod_{i=1}^{k} (x_i!)} \prod_{i=1}^{k} (p_i^{x_i}) \tag{A5}$$

where the summation is again over all vectors $\vec{x} = \{x_1, x_2, \cdots, x_k\}$ satisfying (A1) and (A4). If the k sets are not exhaustive then we define S_{k+1} as above and the same expression (A5) is obtained with k replaced by $k + 1$ in (A1) and (A5) but *not* in (A4), i.e., no condition is applied to the $(k + 1)$th variable.

It is interesting to note that the results for the infinite case ($N = \infty$) can be obtained from those of the finite case by letting N tend to infinity. Table V illustrates this numerically since the four entries in each set correspond to $N = 60$, 120, 360 and ∞, respectively.

APPENDIX II

Approximate Solutions — Infinite and Finite Populations

Let x_i denote the random integral number of observations in a sample of size n which fall in the ith cell ($i = 1, 2, \cdots, k$). If we let

$y_i = x_i - (n/k)$, then the two conditions $\sum_{i=1}^k x_i = n$ and

$$\sum_{i=1}^k y_i = 0 \tag{A6}$$

are equivalent. Let $[x]$ denote the largest integer not greater than x. We shall consider only the case of the equi-probable exhaustive sets.

In the case of an infinite population we wish to compute

$$P = P\left\{n\left(\frac{1}{k} - \beta_i^*\right) \leqq x_i \leqq n\left(\frac{1}{k} + \beta_i^*\right) \quad (i = 1, 2, \cdots, k) \,\middle|\, \sum_{i=1}^k x_i = n\right\}. \tag{A7}$$

If we introduce a continuity correction and use (A6) then we obtain

$$P = P\{-b_i \leqq y_i \leqq a_i (i = 1, 2, \cdots, k) \mid \textstyle\sum_{i=1}^k y_i = 0\} \tag{A8}$$

where for each $i (i = 1, 2, \cdots, k)$

$$a_i = \frac{1}{2} + \left[n\beta_i^* + \frac{n}{k}\right] - \frac{n}{k} \quad \text{and} \quad b_i = \frac{1}{2} + \left[n\beta_i^* - \frac{n}{k}\right] + \frac{n}{k}. \tag{A9}$$

If n/k is an integer and β^* is the common value of $\beta_i^* (i = 1, 2, \cdots, k)$ then $a_1 = a_2 = \cdots = a_k = b_1 = b_2 = \cdots = b_k = a$ (say) and (A8) reduces to

$$P = P\{\,|y_i| \leqq a (i = 1, 2, \cdots, k) \mid \textstyle\sum_{i=1}^k y_i = 0\} \tag{A10}$$

where $a = \frac{1}{2} + [n\beta^*]$.

To compute (A10) *two approximations* are made. The k-variate multinomial probability is first transformed by an orthogonal transformation into a $(k - 1)$-variate distribution with homoscedastic and uncorrelated variables and the *first approximation* is to replace the latter distribution by a multivariate normal distribution with independent variables. The region of integration is the *intersection* of the hypercube $|y_i| \leqq a$ centered at the origin with edge-length $2a$ *and* the hyperplane (A6); the orthogonal transformation merely rotates this intersection about the origin. These intersections are convex figures symmetric with respect to the origin; for example, it is a regular centered hexagon for $k = 3$. These intersections, called Stott figures, are discussed in Appendix III. The *second approximation* made in computing (A10) was to replace the Stott figure by a $(k - 1)$-dimensional central sphere whose radius R is determined by equating the two hypervolumes. Values of R for $k = 2(1)12$ for any a are given in Table IX.

TABLE IX

Intersection $\mathcal{J}$ of the hypercube of edge-length $2a$ centered at the origin and the hyperplane $x_1 + x_2 + \cdots + x_k = 0$.

Dimension k of hypercube	$J(k)$ = Number of equally large simplices in $\mathcal{J}$	Radius R of sphere with content equal to that of $\mathcal{J}$
2	1	1.4142 a
3	6	1.2861 a
4	4	1.3655 a
5	230	1.4436 a
6	66	1.5225 a
7	23,548	1.5995 a
8	2,416	1.6733 a
9	4,675,014	1.7443 a
10	156,190	1.8126 a
11	1,527,092,468	1.8786 a
12	15,724,248	1.9422 a

The content $I(k)$ of $\mathcal{J}$ for all k is given by

$$I(k) = \frac{a^{k-1}\sqrt{k}}{(k-1)!}\left[\binom{k}{0}(k)^{k-1} - \binom{k}{1}(k-2)^{k-1} + \binom{k}{2}(k-4)^{k-1} - \cdots\right]$$

where the terms continue only as long as the arguments k, $k - 2$, $\cdots$ are positive. The radius R of a $(k - 1)$-dimensional sphere of equal content is obtained by equating $I(k)$ and $(R\sqrt{\pi})^{k-1} \Big/ \Gamma\left(\frac{k+1}{2}\right)$.

The orthogonal transformation referred to above is

$$y_i' = \frac{1}{\sqrt{i(i+1)}}(y_1 + y_2 + \cdots + y_i - iy_{i+1}) \qquad (i = 1, 2, \cdots, k) \tag{A11}$$

where y_{k+1} is defined to be identically zero. Then y_k' is identically zero by (A6). The remaining y_i' all have a common variance $\frac{n}{k}$ since for each $i(i = 1, 2, \cdots, k - 1)$

$$\sigma_{y_i'}^2 = \frac{1}{i(i+1)}\left\{i(i+1)n\left(\frac{k-1}{k^2}\right) + 2\binom{i}{2}\left(-\frac{n}{k^2}\right) - 2i^2\left(-\frac{n}{k^2}\right)\right\} = \frac{n}{k} \tag{A12}$$

and are pairwise uncorrelated since for $i < j$

$$\sigma_{y_j'y_i'} = \sigma_{y_i'y_j'} = \frac{1}{\sqrt{i(i+1)j(j+1)}} \left\{ \frac{ni(k-1)}{k^2} + 2\binom{i}{2}\left(-\frac{n}{k^2}\right) + i(j-i)\left(-\frac{n}{k^2}\right) - i(j-1) \left(-\frac{n}{k^2}\right) - \frac{ni(k-1)}{k^2} + (ij-ij)\left(-\frac{n}{k^2}\right)\right\} = 0. \tag{A13}$$

If we let $\nu = k - 1$, let $r = R/\sigma = R\sqrt{k/n}$ and let S denote the central sphere of radius r then the approximate probability (dropping primes) is given by

$$P = \int_S \cdots \int \left(\frac{1}{2\pi}\right)^{\nu/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{\nu} y_i^2\right\} dy_1\, dy_2 \cdots dy_\nu \tag{A14}$$
$$= P\,\{\chi_\nu^2 \leqq r^2\}$$

where χ_ν^2 denotes a chi-square random variable with ν degrees of freedom.

In the case of a finite population of size N the only change in the above discussion is to replace (A12) by

$$\sigma_{y_i'}^2 = \frac{n}{k}\left(\frac{N-n}{N-1}\right) \quad (i = 1, 2, \cdots, k-1) \tag{A15}$$

thus increasing the value of r^2 and the value of P; this decreases n if P is held fixed at any P^*. If we let n_N and n_∞ denote the required values for a finite population of size N and an infinite population, respectively, for the same fixed k, β^* and P^* then we obtain from (A14) and (A15)

$$n_\infty \cong n_N\left(\frac{N-n_N}{N-1}\right), \tag{A16}$$

or, taking the smaller solution in n_N, we have for large N

$$n_N \cong \frac{N - \sqrt{N^2 - 4(N-1)n_\infty}}{2}. \tag{A17}$$

Replacing $N - 1$ by N in (A16) we easily obtain for large N the simpler result

$$\frac{1}{n_N} \cong \frac{1}{n_\infty} - \frac{1}{N}. \tag{A18}$$

The error in P involved in both of the above approximations (A14) and (A17) is evaluated in Table VII for $N = 120$ and $N = \infty$ for selected values of n, β^* and k.

If n/k is not an integer then the above discussion may not apply since

a_i may not equal b_i in (A9). Assuming again a common β^* then we have a common "a" and a common "b" in (A9). In this case, averaging the approximate probabilities obtained by using $2a$ and $2b$ alternately as the edge-length of the hypercube was found to be satisfactory for computing the tables of this paper.

APPENDIX III

Geometric Results and Eulerian (Diamond) Numbers

The problem here is to find the $(k-1)$-dimensional content (or hypervolume) of the intersection $\mathcal{J}$ of the centered k-dimensional hypercube $|y_i| < a(i = 1, 2, \cdots, k)$ and the $(k-1)$-dimensional hyperplane $y_1 + y_2 + \cdots + y_k = 0$. The geometry for even k and odd k is quite different. The number of vertices of $\mathcal{J}$ for even k and odd k, respectively, is

$$\binom{k}{k/2} \quad \text{and} \quad k\binom{k-1}{(k-1)/2}; \tag{A19}$$

for example, for $k = 3$ we obtain the $3\binom{2}{1} = 6$ vertices $(a, -a, 0)$, $(-a, a, 0)$, $(a, 0, -a)$, $(-a, 0, a)$, $(0, a, -a)$ and $(0, -a, a)$. The vertices are all equally distant from the origin. All the edges of $\mathcal{J}$ have a common length $d = d(k)$ which equals $2a\sqrt{2}$ for even k and $a\sqrt{2}$ for odd k. The intersection $\mathcal{J}$ is a convex figure which is symmetric with respect to the origin and is known as a Stott figure.[6] The Stott figure can be partitioned into an integral number $J(k)$ of $(k-1)$-dimensional simplices which are not necessarily regular but are such that each simplex has the same content as a regular $(k-1)$-dimensional simplex with edge-length d. Hence, using a result on page 125 of Reference 8, the content $I(k)$ of $\mathcal{J}$ is given by

$$I(k) = \left(\frac{d\sqrt{2}}{2}\right)^{k-1} \frac{\sqrt{k}}{(k-1)!} J(k). \tag{A20}$$

The integers $J(k)$ are given in the middle column of Table IX; for example, the integer 6 for $k = 3$ indicates that there are six equilateral triangles in the centered hexagon.

D. Slepian[7] has shown that for even k the integers $J(k)$ can be found by generating a "triangle" of numbers using the recurrence relation

$$S_{i,j} = jS_{i-1,j} + iS_{i,j-1} \quad (i, j = 1, 2, \cdots) \tag{A21}$$

with boundary conditions $S_{1,j} = S_{j,1} = 1$ for all j; then the desired

quantities are

$$S_{i,i} = J(2i) \qquad (i = 1, 2, \cdots). \quad \text{(A22)}$$

Similarly for odd k he showed that we can use the recurrence relation

$$T_{i,j} = (2j + 1)T_{i-1,j} + (2i + 1)T_{i,j-1} \quad (i, j = 1, 2, \cdots) \quad \text{(A23)}$$

with boundary conditions $T_{0,j} = T_{j,0} = 1$ for all j; then the desired quantities are

$$T_{i,i} = J(2i + 1) \quad (i = 1, 2, \cdots). \qquad \text{(A24)}$$

Fig. 9 shows these numbers in two diamond-shaped patterns and explains another interesting way of obtaining these numbers.

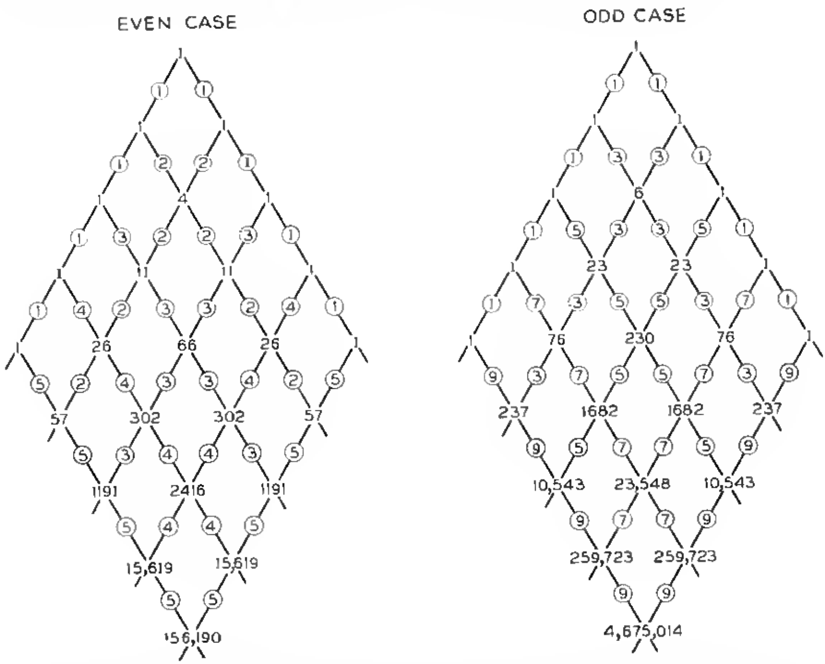

Fig. 9 — Combinatoric derivation of certain Eulerian (diamond) numbers. The number at any vertex V is obtained by considering any one path from the top vertex to V, multiplying the circled numbers encountered in this path, and summing the results obtained over all possible downward paths from the top vertex to V. In particular, the values on the vertical diagonal (of the diamond) are the values of $J(k)$ in Table IX. It is interesting to note that the sum of all the uncircled numbers in the mth row is $2^{m-1}(m-1)!$ for the odd case and $m!$ for the even case. This is shown above for $m = 1, 2, 3, 4, 5$ and would hold for all m if this pattern were continued indefinitely. The *circled* numbers are obtained by numbering the parallel diagonal lines starting with one at the "top," using all positive integers in the even case and only odd integers in the odd case.

The integers $J(k)$ arise in connection with combinatorial problems. As an example for even k, suppose we draw at random m balls in succession from an urn containing m balls marked $1, 2, \cdots, m$. Let X denote the number of times that the observed number increases, (say) always counting the first draw as an increase. Then it can be shown that

$$P\{X = j\} = S_{j,m+1-j}/m! \quad (j = 1, 2, \cdots, m), \tag{A25}$$

i.e., the mth row of the left diamond Fig. 9 divided by the sum $m!$ of that row gives the elementary probability distribution of X.

The problem of computing (A25) also arose in the work of V. H. Moore and W. A. Wallis[4] and M. MacMahon[3] who referred to it as Simon Newcomb's problem. J. Riordan[5] has studied the numbers $J(k)$ for even k and Carlitz and Riordan[6] call them Eulerian numbers (to be distinguished from the classical Euler numbers); an explicit formula as well as a generating function appears in these papers. The $S_{i,j}$ are related to the Eulerian numbers $A_{n,k}$ (defined in Reference 5) by $S_{i,j} = A_{i+j-1,j}$.

Explicit expressions for $J(k)$ for odd and even k are obtainable from (A22), (A24) and the more general results

$$S_{i,j} = \sum_{\alpha=0}^{j-1} (-1)^\alpha \binom{i+j}{\alpha} (j - \alpha)^{i+j-1} \tag{A26}$$

$$T_{i,j} = \sum_{\alpha=0}^{j} (-1)^\alpha \binom{i+j+1}{\alpha} [2(j - \alpha) + 1]^{i+j} \tag{A27}$$

due to D. Slepian.[7] It is easily shown that these formulae satisfy the corresponding recurrence relations as well as the boundary conditions. By an induction and symmetry argument applied to (A21) and (A23) and the boundary conditions it is easy to prove that

$$S_{i,j} = S_{j,i} \quad \text{and} \quad T_{i,j} = T_{j,i}. \tag{A28}$$

Substituting (A26) and (A27) in (A28) gives rise to interesting, nontrivial identities. For completeness we also give the generating functions derived by D. Slepian[7]

$$\sum_{i,j=1}^{\infty} \frac{S_{i,j} t^i u^j}{(i + j - 1)!} = \frac{tu(e^t - e^u)}{te^u - ue^t} \tag{A29}$$

$$\sum_{i,j=1}^{\infty} \frac{S_{i,j} t^i u^j}{(i + j)!} = \log_e \left[\frac{t - u}{te^u - ue^t} \right] \tag{A30}$$

$$\sum_{i,j=0}^{\infty} \frac{T_{i,j} t^i u^j}{(i + j)!} = \frac{(t - u)e^{t+u}}{te^{2u} - ue^{2t}}. \tag{A31}$$

The final result for the content $I(k)$ of $\mathcal{S}$ can, using the above be written as a *single* expression

$$I(k) = a^{k-1} \frac{\sqrt{k}}{(k-1)!} \sum_{\alpha=0}^{[(k-1)/2]} (-1)^{\alpha} \binom{k}{\alpha} (k - 2\alpha)^{k-1} \qquad \text{(A32)}$$

for all k where $[x]$ denotes the largest integer not greater than x. It has been pointed out by J. W. Tukey that (A32) can also be obtained by probabilistic considerations and that it appears in Laplace's "Theorie Analytique" (Book 2, page 260).

APPENDIX IV

Remarks on the Confidence Bands

It should be remarked that other assumptions on the true, unknown distribution can be used in conjunction with the confidence bands obtained in Section VI. It has been pointed out by J. W. Tukey, for example, that in the case of the first diagram in Fig. 8 the experimenter might be willing to assume that the true distribution is unimodal and that the mode x_m is such that $x_m \leqq x_{64}$. Then on purely geometrical considerations it can be shown that the confidence band can be modified as shown in the first diagram of Fig. 10. Briefly, if the true distribution enters any one of the three deleted triangles with any slope s then in order to get out again without leaving the confidence band the slope must get larger than s. But this contradicts the assumption that the density steadily decreases after x_{64}.

Similarly, with the same problem, if the experimenter assumes that the true distribution is unimodal and that $x_{73} \leqq x_m \leqq x_{88}$ then the first diagram of Fig. 8 can be modified as in the second diagram of Fig. 10. The assumption of unimodality is reasonable in many different practical applications but has not often been utilized in statistical techniques.

It is possible to formulate a problem for fixed P^* and n which requires the determination of that k which makes the *maximum* (or some average) *vertical width* of the confidence bands as small as possible. For example, for $P^* = 0.85$ and $n = 240$ the value $k = 10$ minimizes the maximum vertical width. It should be pointed out that if the experimenter's principal interest is in finding confidence bands with small vertical widths then this procedure appears to be quite inefficient compared with that based on the Kolmogorov statistic.[1]

A proper comparison is difficult since the nominal P^* is a lower bound and not the correct value of the confidence level associated with the pro-

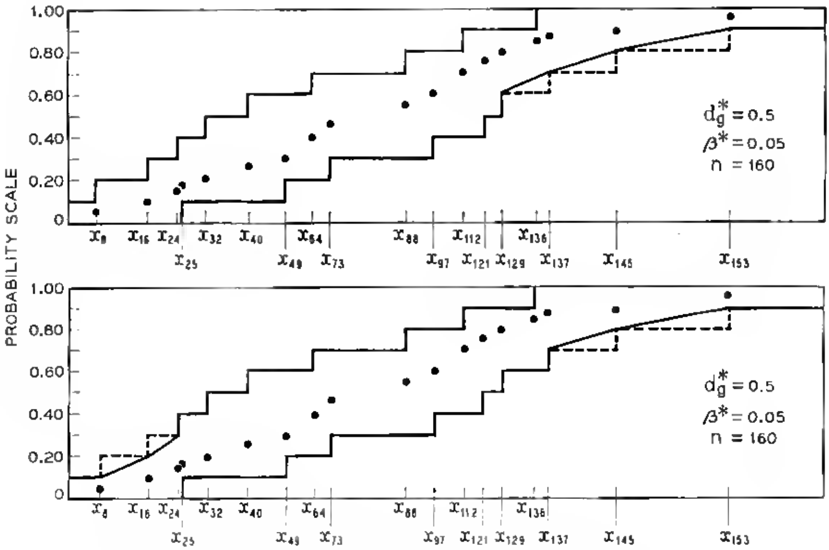

Fig. 10 — Modified confidence bands which include the true distribution function with confidence greater than $P^* = 0.85$ for $k = 10$ and $d_g^* = 0.5$.

posed confidence bands. As mentioned in the body of the paper the development of a confidence band is just a by-product of the main theme of this paper which is the representativeness of the sample.

VII. CONCLUSION

Definitions of representativness and of degree of representativeness are given and tables are included which give the sample size required to guarantee with preassigned probability P^* that a random sample will satisfy a condition of representativeness, the definition of which is agreed upon in advance. Thus, for experimenters who wish to know *in advance* how many observations will be needed for a distribution study, tbe problem has been given a precise nonparametric formulation and the solution has been found for some cases.

This formulation also leads to confidence bounds on the unknown distribution *after the observations are taken*. Examples are given to illustrate this.

The tables for the case of pairwise disjoint, equi-probable and exhaustive cells may also prove to be useful for the problem of determining the sample size required to obtain *simultaneous* confidence limits (on a preassigned level P^*) for *all* of the cell probabilities of a multinomial

distribution. Further investigation is needed to state precisely the conditions under which these tables can be used for this related problem.

VIII. ACKNOWLEDGEMENT

The authors wish to acknowledge the encouragement of the Bell Laboratories staff at Allentown and various helpful suggestions and aid received from D. Slepian, S. Monro, J. W. Tukey, J. Riordan, T. W. Anderson and others. The assistance of Miss Phyllis Groll in checking tables and proofreading is gratefully acknowledged.

REFERENCES

1. Birnbaum, Z. W., Numerical Tabulation of the Distribution of Kolmogorov's Statistic for Finite Sample Size, Journal of the American Statistical Association, **47,** Sept., 1952.
2. Loève, M., Probability Theory, D. Van Nostrand Company, New York, 1955.
3. MacMahon, P. A. *Combinatory Analysis*, Vol. 1, p. 187, Cambridge, 1915.
4. Moore, V. H. and Wallis, W. A., Time Series Significance Tests Based on Signs of Differences, Journal of the American Statistical Association, **38,** June, 1943.
5. Riordan, J., Triangular Permutation Numbers, Proc. of the Amer. Math. Soc., **2,** June, 1951; Carlitz, L. and Riordan, J., Congruences for Eulerian Numbers, Duke Math. J., **20,** September, 1953.
6. Schoute, Analytical Treatment of the Polytopes Regularly Derived from the Regular Polytopes (IV), Verhandelingen der Koninklijke Akademie van Wetenschappen te Amsterdam (eerste sectie), **11.5,** pp. 73–108, 1913.
7. Slepian, D., On the Volume of Certain Polytopes, internal B.T.L. memorandum, April 11, 1956.
8. Somerville, D.M.Y., *An Introduction to the Geometry of N Dimensions*, p. 125 E. P. Dutton and Company, N. Y., 1929.